Interacción emocional con actores virtuales a través de lenguaje natural

Sandra Baldassarri, Eva Cerezo, David Anaya
Grupo de Informática Gráfica Avanzada (GIGA)
Dept. Informática e Ingeniería de Sistemas
Universidad de Zaragoza
Instituto de Investigación en Ingeniería de Aragón (I3A)
Zaragoza
sandra,ecerezo,anaya@unizar.es

Resumen

Las nuevas tendencias en Interacción Hombre Máquina se centran en la utilización de técnicas que permiten que el usuario se comunique y, por tanto, se relacione, con las máquinas de forma natural. De entre estas técnicas, el lenguaje natural desempeña un papel fundamental. Sin embargo, por sí solo el lenguaje no es suficiente, y es necesario que se añadan aspectos emocionales al hablar para mejorar la comunicación. Este artículo describe una interfaz conversacional que soporta la comunicación del usuario con un actor virtual en tiempo real, y utilizando lenguaje natural y emocional en español. El estado emocional del actor virtual puede cambiar a lo largo del desarrollo de la conversación con un usuario y sus emociones se pueden expresar variando las respuestas y modulando la voz con la emoción adecuada.

1. Introducción

El alcance actual del uso de tecnologías de la información ha producido un creciente interés en aplicaciones y sistemas informáticos que no requieren un conocimiento tecnológico previo. Este hecho ha provocado que se mejoren las técnicas de Interacción Hombre Máquina (IHM) para permitir una comunicación natural y personalizada entre usuarios y máquinas. Uno de los métodos de comunicación más comunes es el lenguaje natural. Sin embargo, y a pesar de que el método ha sido ampliamente estudiado, aún permanecen varios aspectos relevantes sin resolver. Para obtener una interacción más natural y más creíble, los sistemas de IHM deben ser capaces de responder apropiadamente a los usuarios por medio de una reacción afectiva [16]. Dentro de la comunicación verbal las reacciones afectivas se logran añadiendo variabilidad en las respuestas y mediante la introducción de emociones en la síntesis del habla [7].

La incorporación de emociones en la voz se realiza mediante cambios en las estructuras melódicas y rítmicas [4]. En los últimos años, se han desarrollado varios trabajos que consideran las componentes emocionales en la síntesis de la voz, sin embargo, la mayoría de los estudios en este área se refieren al idioma inglés [15] [19] [10]. En el caso del idioma español, el trabajo de Montero *et al* [14] se centra en el análisis prosódico del español y el modelado de cuatro emociones en una base de datos. En dicho trabajo se hace un experimento interesante sobre la relevancia de la cualidad de la voz en el reconocimiento de estados emocionales. En cambio, Iriondo *et al* [11] presentan un conjunto de reglas que describen el comportamiento de los parámetros más significativos del habla relacionados con expresiones emocionales y validan el modelo utilizando técnicas de síntesis del habla, y simulando siete emociones básicas. Un estudio similar fue hecho en [5], pero obteniendo las expresiones emocionales a partir de vídeos realizados por actores profesionales tanto en inglés, como en francés y en español.

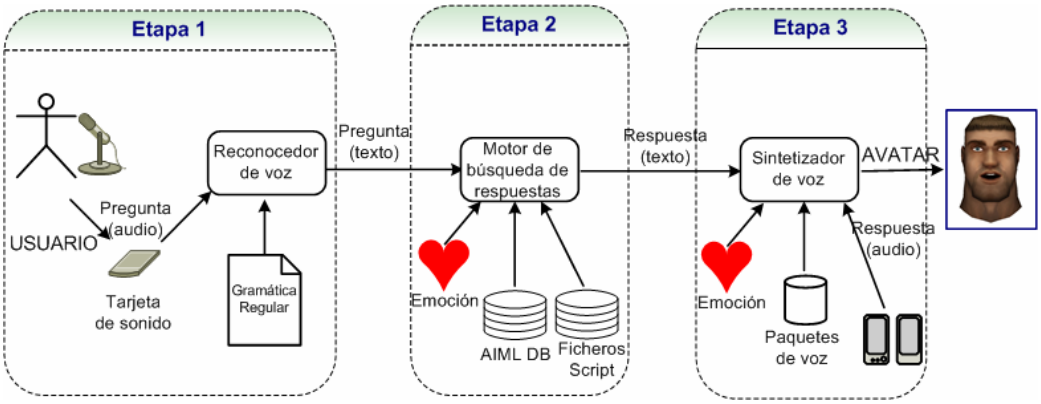

Figura 1. Etapas del proceso de comunicación por voz usuario-avatar.

En este artículo se presenta un sistema para la generación de síntesis de voz emocional en español, pero a diferencia de los trabajos anteriores permite la *interacción* con el usuario en lenguaje natural y soporta la comunicación en *tiempo real*. Además, la interfaz conversacional desarrollada considera el estado de ánimo del actor virtual, que puede variar dependiendo de la relación con el usuario a lo largo de la conversación, y que se expresa mediante la adecuada modulación de la voz y mediante la selección de la respuesta correcta para el usuario dependiendo de ese estado. Para lograr este propósito el sistema guarda información acerca de la "historia" de la conversación.

A continuación, en la sección 2, se presenta el trabajo desarrollado en interacción emocional a través de lenguaje natural, dando, en primer lugar una visión general de las etapas que forman el proceso de comunicación entre un usuario y un avatar, explicando el modo en que se ha llevado a cabo la gestión emocional, y presentando un ejemplo de conversación con actores virtuales. La sección 3 detalla los resultados de la generación de voz con emoción. Mientras que la sección 4 refleja el rendimiento del sistema desarrollado. Por último, en la sección 5, se comentan las conclusiones y las líneas de trabajo futuro.

2. Interacción emocional a través de lenguaje natural

2.1. Visión general del proceso de comunicación usuario-avatar

El proceso global de comunicación entre un usuario y un actor virtual a través de la voz se realiza en tres etapas, tal y como se muestra en la Figura 1. En los siguientes apartados se presenta una explicación detallada de cada una de las tres etapas.

Etapa 1: Reconocimiento de la voz.
El objetivo de esta primera etapa consiste en obtener una cadena de texto a partir de las palabras pronunciadas por el usuario en español. Para lograr este objetivo se ha desarrollado un reconocedor de voz a partir del software Loquendo ASR (Audio Speech Recognition) [12]. En base a la biblioteca dinámica del ASR se ha construido un dispositivo de reconocimiento que permite:

- Capturar el audio y darle un formato adecuado para ser procesado posteriormente por el reconocedor
- Extraer palabras del audio por medio de gramáticas regulares que indican las palabras que se pueden reconocer
- Procesar los resultados devueltos por la función de reconocimiento para saber si no se ha reconocido algo, si se ha escuchado ruido o si el usuario debe hablar más claro o más rápido
- Establecer un modo de funcionamiento síncrono o asíncrono pudiendo adaptar su uso a diferentes aplicaciones

- Permitir reconocer palabras habladas en diferentes idiomas

Para obtener una respuesta precisa dentro de un periodo de tiempo razonable Loquendo ASR permite trabajar con tres posibles tipos de gramáticas independientes de contexto: ABNF (Augmented BNF), XMLF (XML Form) y JSGF (Java Speech Grammar Format). De ellas, en este trabajo se ha optado por JSGF ya que su sintaxis evita el farragoso etiquetado del XMLF y porque es utilizada por una comunidad superior a la de ABNF.

Debido a la necesidad de que el sistema fuera capaz de "entender" y hablar español, durante el desarrollo del reconocedor se tuvieron que resolver ciertos problemas específicos provenientes de la implementación para dicho idioma. En particular, Loquendo ASR no es capaz de distinguir entre palabras con o sin 'h', con 'b' o 'v', o con 'y' o 'll'. Tampoco es capaz de detectar la diferencia entre palabras que se escriben igual pero se pronuncian diferente, donde la diferencia radica en un acento, como por ejemplo: la forma verbal "está" y el pronombre "ésta". Para solucionar estos problemas se toma la decisión de escribir todas las palabras de la gramática sin tildes y para el caso de aquellas que se pronuncian igual pero se diferencian en que se escriben con o sin 'h', 'b' o 'v', 'll' o 'y', escribirlas siempre con 'h', 'b' y 'll'. Con este cambio se resuelven los problemas, pero en ningún momento se disminuye la eficacia del sistema. Sin embargo, este hecho se debe tener en cuenta a la hora de escribir las palabras de la gramática y a la hora de generar los ficheros con las posibles respuestas correctas.

Etapa 2: Obteniendo la respuesta correcta

Esta etapa se encarga básicamente de generar una respuesta a las preguntas del usuario, que han sido obtenidas en modo texto de la etapa anterior (ver Figura 1). El motor de búsqueda desarrollado se basa en el proyecto CyN [17], basado a su vez en ALICE [1] tanto en su motor como en las directrices y normas que especifica. No obstante, CyN está diseñado únicamente para mantener conversaciones en inglés, así que fue necesario modificar el código para poder soportar los diálogos en español. Las principales diferencias radican en ser capaz de trabajar con acentos, diéresis, eñes y permitir caracteres de apertura de interrogación o exclamación.

El proceso de búsqueda de respuestas se basa en el reconocimiento de patrones de palabras en la pregunta del usuario, donde cada patrón lleva asociadas, como mínimo, respuestas fijas que conforman el conocimiento estático del sistema. Sin embargo, las respuestas pueden experimentar variaciones aleatorias para que el usuario no tenga la impresión de repetición si la conversación se alarga, o las respuestas pueden diferir en función del estado de ánimo del actor virtual, formando así el conocimiento dinámico. En nuestro caso, este tipo de respuestas del sistema es suficiente, ya que los temas de las conversaciones son, de momento, bastante específicos (entornos educativos u órdenes específicas para manejar entornos domóticos).

El conocimiento estático y dinámico del personaje virtual se especifica en AIML (Artificial Intelligence Markup Language) [2]. AIML es un derivado del XML, cuyo poder radica en varios aspectos básicos:

- Su sintaxis permite extraer fácilmente el contenido semántico de una pregunta para poder devolver la respuesta adecuada rápidamente.
- Permite utilizar etiquetas para combinar las respuestas, aumentando de esa forma la variedad de éstas y el número de preguntas a las que se puede dar una contestación.
- Permite utilizar la recursividad para dar una respuesta a una pregunta dada para la cual, en teoría, no hay una respuesta directa.
- Permite llevar un historial de la conversación, lo que facilita la detección del estado emocional del hablante en su relación con el avatar y cambiar el estado de ánimo del actor virtual y sus respuestas en consecuencia.

El intérprete AIML ha sido modificado para incluir comandos o llamadas a ficheros script dentro de las categorías AIML. Estos comandos se ejecutan cuando se activa la categoría en la que están declarados y cuando se devuelve su resultado como parte de la respuesta al usuario. Esto hace posible, por ejemplo, consultar la hora del sistema, conectarse a una página web para consultar la temperatura de cualquier ciudad, cambiar el estado emocional del avatar, etc.

```
<category>
    <pattern>
       CREO QUE DEBERIAMOS DEJAR ESTA CONVERSACION
    </pattern>
    <template>
       <random>
             <li>
                 <sad/>Bueno, supongo que no soy lo que
                 esperabas
             </li>
             <li>
                 <angry/>¿Pero qué te pasa?¿es que no te gusto?
             </li>
              <li>
                <surprised/>¿Por qué? Con lo bien que lo
                 estábamos pasando
              </li>
              <li>
                 <happy/>Venga, no seas soso,
                 Sigamos un poco más
              </li>
              <li>Vale, seguiremos en otro momento
               </li>
        </random>
     </template>
 </category>
```





Figura 2. Ejemplo de una categoría AIML en la que la respuesta es dependiente del estado emocional

Etapa 3: Conversión texto a audio.
La síntesis de la voz se realiza por medio de SAPI5 [13], un conjunto de bibliotecas y funciones que permiten implementar un sintetizador de voz en inglés y se puede descargar y utilizar de forma gratuita, pero en este proceso se utilizan los paquetes de Loquendo para generar voz en castellano. SAPI5 permite obtener la información acerca de los visemas (representación visual de un fonema) que se producen al pronunciar la frase que se quiere sintetizar. La información de estos visemas se utiliza para modelar los movimientos de la boca del avatar al hablar, de esta manera se consigue la sincronización labial del avatar con la frase que se está sintetizando. Para evitar que la voz suene artificial, ésta ha sido dotada con una componente emocional. La generación de emociones en tiempo real se realiza modificando la modulación de la voz trabajando sobre los parámetros de tono, escala de la frecuencia, volumen y velocidad del habla, que posee SAPI5 (ya sea a través de funciones globales como a través de etiquetas XML introducidas dentro de la cadena de texto que se va a sintetizar), tal y como se detalla en la siguiente sección.

2.2. Gestión emocional

En nuestro sistema se ha optado por trabajar con las seis emociones universales de Ekman [8]: alegría, tristeza, enfado, sorpresa, aversión y miedo, más la neutral. Dichas emociones no sólo se tienen en cuenta en la síntesis de voz (modulación) sino también en la generación de las respuestas en dos niveles:

1. La respuesta depende del estado emocional del actor virtual. Por esta razón, se ha rediseñado el comando AIML <random> para añadir esta característica. Como se puede ver en ejemplo de la Figura 2 puede haber más de una respuesta con la misma etiqueta: <angry>, <sad>, … en este caso, una de estas respuestas será devuelta de forma aleatoria. Siempre debe

Figura 3. Un usuario hablando con el avatar.

haber una respuesta que no tiene una etiqueta asociada y es la que corresponde al estado emocional "neutral".

2. El estado emocional del actor se puede modificar a lo largo de una conversación, dependiendo del desarrollo de la misma. Por ejemplo, si se está hablando sobre temas que le gustan se alegra, si se le da información que no conocía se asombra, si se le amenaza se asusta, si el usuario insulta al avatar, éste se enfada,…

Determinadas palabras usadas en las respuestas o preguntas del usuario sirven de clave para identificar el estado de ánimo del usuario y modificar el estado del avatar en consecuencia. Las categorías AIML que se activan con estas palabras clave contienen comandos dentro de la respuesta, que se ejecutan internamente para modificar el estado de ánimo del avatar. En la siguiente categoría se muestra cómo al insultar al avatar éste cambia su estado de ánimo y pasa a estar enfadado:

```
<category>
 <pattern> ERES TONTO </pattern>
 <template>
  <script>emotion=angry</script>
  Yo no te he insultado
 </template>
</category>
```

De este modo, cuando el usuario le dice al avatar "Eres tonto" se activa esta categoría. En ella se ejecuta en primer lugar el comando `script` que asigna el valor "angry" a la variable que controla el estado de ánimo del avatar. Después, el avatar contesta "Yo no te insultado", sintetizando la frase con los valores del tono, velocidad y volumen asociados al enfado.

2.3. Ejemplo: Conversación simple con actores virtuales

La interfaz conversacional desarrollada permite al usuario mantener una conversación en español con un actor virtual, tal y como se puede ver en la Figura 3. Dicha interfaz ha sido añadida a una aplicación, denominada MaxinePPT [18], que genera presentaciones 3D realizadas por un personaje virtual a partir de diapositivas de un PowerPoint. El nuevo sistema permite la interacción en lenguaje natural (y en castellano) entre el usuario y el presentador virtual, de tal forma que el usuario puede preguntar cuestiones acerca de la presentación y obtener respuestas relacionadas con el tema. De hecho, esta interfaz ha sido utilizada como plataforma para simplificar y mejorar la enseñanza y la presentación de temas de Informática Gráfica en la carrera de Ingeniería Informática [6].
Además, el sistema también está preparado para recibir comandos, funcionalidad que actualmente se está desarrollando en otra aplicación donde un actor virtual ayuda a los usuarios en el manejo de un sistema domótico.

3. Resultados: Generación de voz con emoción.

La voz generada por los conversores texto-voz normalmente suena artificial, siendo ésta una de las causas de rechazo de los avatares por

parte del público general. La introducción de las emociones en un sistema de síntesis de voz resulta fundamental para conseguir naturalidad y credibilidad en la interacción por voz.

El modelado de las distintas emociones se realiza asignando unos valores fijos a los parámetros de tono, volumen y velocidad del habla que proporciona SAPI5. La configuración de estos parámetros se basa en los estudios de Boula et al [5], Francisco et al [9], Iriondo et al [11] y Barra et al [3].

Para encontrar los valores específicos a los que se deben fijar estos parámetros con el objetivo de obtener cada una de las seis emociones se decidió realizar una encuesta de valoración. El proceso se llevó a cabo con 15 personas de diferente edad y sexo, que tuvieron que escuchar el resultado de la síntesis de diferentes frases, modeladas y sintetizadas con distintas emociones. Para recoger sus impresiones y poder determinar la credibilidad de las emociones sintetizadas, se utilizaron los tres métodos de evaluación que se detallan a continuación:

1. Elección forzada: Este método consiste en facilitar a los sujetos el conjunto finito de las posibles respuestas, que en este caso engloban las 6 emociones modeladas más la neutral. Las ventajas de este método son: que es fácil de llevar a cabo, que proporciona una medida simple de reconocimiento y que permite comparar distintos estudios. Sin embargo, tiene una gran desventaja y es que no proporciona información acerca de la calidad del estímulo desde el punto de vista de la naturalidad y la veracidad.
2. Elección libre: En este método la respuesta no se restringe a un conjunto cerrado de emociones. Los sujetos nombran la emoción que ellos consideran que se estaba sintetizando pero sin saber cuales son las que se pretenden conseguir. Este modelo está especialmente indicado para encontrar fenómenos inesperados durante el experimento.
3. Elección libre modificada: En este método, dependiendo de la emoción que se está probando, se emplean dos frases diferentes: una frase neutra, que no tenga ninguna connotación emocional (p.e. “mi casa es azul”) y una frase que pueda evocar esa emoción o que pueda ser usada por una persona que tenga ese estado de ánimo (p. e. para la emoción “sorpresa” la frase podría ser “No me lo puedo creer”). Como medida del impacto de la prosodia en la percepción se considera la diferencia de los resultados obtenidos entre el reconocimiento de la emoción sintetizada. Es decir, primero siempre se sintetiza la frase neutra con la prosodia de la emoción. El sujeto de la prueba nombrará la emoción que le sugiere. A continuación se sintetiza el texto con emoción y se anota la respuesta del sujeto. Sólo se toma como resultado satisfactorio si tanto la síntesis de la frase con connotación y la frase neutra han conseguido evocar al sujeto de la prueba la emoción testeada.

Una vez realizadas estas tres pruebas se corrigen y se ajustan los valores de los tres parámetros que controlan la síntesis emocional y se vuelven a realizar las pruebas de evaluación. Este proceso se repite hasta que los resultados de las pruebas son satisfactorios. Las Tablas 1, 2 y 3 muestran los resultados obtenidos de los *tests* de evaluación con los valores que finalmente se han aprobado como satisfactorios. En todas las tablas, la primera columna de la izquierda indica la emoción que se está sintetizando y la fila superior indica las emociones que les ha sugerido a los participantes. La última columna de las Tablas 2 y 3 se refiere a “Otras” emociones diferentes de las 6 emociones sintetizadas.

Los valores validados por los *tests* de los parámetros que permiten modular la síntesis de la voz para generar emociones (tono, volumen y velocidad) se muestran en la Tabla 4. En base a las pruebas realizadas es posible concluir que las emociones mejor sintetizadas son la tristeza, el enfado y la aversión. Sin embargo, y dado que el reconocimiento de emociones es algo subjetivo y difícil, probablemente sería aconsejable realizar un análisis más exhaustivo de los resultados obtenidos, realizando pruebas con un número más amplio de personas.

	Alegría	Aversión	Enfado	Miedo	Neutral	Sorpresa	Tristeza
Alegría	70%	0%	0%	0%	20%	10%	0%
Aversión	0%	80%	20%	0%	0%	0%	0%
Enfado	0%	10%	80%	0%	10%	0%	0%
Miedo	0%	0%	0%	80%	0%	10%	10%
Neutral	0%	0%	0%	0%	100%	0%	0%
Sorpresa	0%	0%	0%	20%	0%	80%	0%
Tristeza	0%	0%	0%	0%	0%	0%	100%

Tabla 1. Elección forzada.

	Alegría	Aversión	Enfado	Miedo	Neutral	Sorpresa	Tristeza	Otras
Alegría	50%	0%	0%	0%	0%	30%	0%	20%
Aversión	0%	80%	20%	0%	0%	0%	0%	0%
Enfado	0%	0%	70%	10%	0%	0%	0%	20%
Miedo	0%	0%	0%	70%	0%	0%	10%	20%
Neutral	0%	0%	0%	0%	90%	0%	0%	10%
Sorpresa	10%	0%	0%	0%	20%	60%	0%	10%
Tristeza	0%	0%	0%	0%	0%	0%	80%	20%

Tabla 2. Elección libre.

	Alegría	Aversión	Enfado	Miedo	Neutral	Sorpresa	Tristeza	Otras
Alegría	30%	0%	0%	0%	40%	0%	0%	30%
Aversión	0%	40%	10%	0%	10%	0%	10%	30%
Enfado	0%	10%	60%	10%	0%	0%	0%	20%
Miedo	10%	0%	0%	40%	0%	0%	20%	30%
Neutral	0%	0%	0%	0%	100%	0%	0%	0%
Sorpresa	0%	10%	0%	10%	20%	40%	0%	20%
Tristeza	0%	10%	0%	0%	0%	0%	60%	30%

Tabla 3. Elección libre modificada.

	Volumen (0 - 100)	Velocidad (-10 - 10)	Tono (-10 - 10)
Alegría	80	3	4
Aversión	50	3	-6
Enfado	70	3	0
Miedo	56	1	2
Neutral	50	0	0
Sorpresa	56	0	3
Tristeza	44	-2	2

Tabla 4. Parámetros establecidos para el volumen, velocidad y tono para la generación de voz con emoción.

Etapas de la conversación	T. Mínimo	T. Máximo	T. Medio
Reconocimiento de voz	1.6 s.	2.01 s.	1.78 s.
Síntesis de voz	0.18 s.	0.2 s.	0.3 s.
Búsqueda de resultados	0.1 s.	0.17 s.	0.2 s.

Tabla 5. Medidas de tiempo de las diferentes etapas de una conversación en segundos.

4. Rendimiento del sistema

Debido a los requerimientos de tiempo real, se ha intentando reducir al mínimo el tiempo que tarda el sistema desde que el usuario termina de hablar hasta que el usuario empieza a escuchar la respuesta. Un tiempo excesivo de espera disminuiría la sensación de interactividad con el sistema y provocaría el desagrado del usuario. Como es lógico, el tiempo de respuesta del sistema varía dependiendo de lo larga que es la consulta del usuario y de lo larga que es la respuesta.

La medición de los tiempos de cada etapa del proceso de comunicación con el avatar, se ha realizado a lo largo de la conversación mantenida entre un usuario y el avatar, con frases cuyo número de palabras varían entre 1, la más corta, y 20, la más larga. Las pruebas de reconocimiento y síntesis de voz se han efectuado en castellano.

Las mediciones de tiempo en la búsqueda de respuestas se han realizado con el cerebro de ALICE [1], que posee unas 47000 categorías. Aunque el contenido de estas categorías está en inglés, este hecho no influye a la hora de medir los tiempos de búsqueda. De otro modo no hubiese sido factible crear tal cantidad de categorías en castellano, de forma que se pudiese tener un sistema cargado de respuestas para hacer las mediciones del tiempo de búsqueda. Para hacer estas pruebas se buscó respuesta a 200 frases diferentes en inglés y se calculó los tiempos de respuesta medio, mínimo y máximo.

La Tabla 5 muestra el estudio de tiempos realizado a lo largo de varias conversaciones que mantuvieron 20 usuarios con el avatar. En esta Tabla, tanto para la síntesis como para el reconocimiento de voz, el tiempo máximo se corresponde con el reconocimiento o síntesis de las frases más largas. Obviamente cuanto más larga es la frase más tardan estos procesos, sin embargo los tiempos que se obtienen en la síntesis son breves en comparación con los tiempos obtenidos del reconocimiento de voz.

La diferencia de tiempo entre reconocer una frase de una palabra y una de 20 palabras es tan solo de 0.41 segundos (2.01s - 1.6s). Esto hace pensar que en el proceso de reconocimiento se realizan unos pasos previos de entorno a 1.4 segundos de duración y que son independientes del tamaño de la frase a reconocer. En general, los tiempos de búsqueda de respuesta son muy buenos, y cabe destacar que el tiempo máximo no corresponde con la frase más larga sino con la que más operaciones recursivas activa (propias del AIML) en el proceso de búsqueda de respuesta cuando se está recorriendo la estructura de árbol que almacena las respuestas. Los tiempos obtenidos de forma global son aceptables, ya que el tiempo medio de las tres etapas juntas es de 2.28 segundos, lo que permite que el ritmo de la conversación sea razonable para la comunicación en tiempo real.

5. Conclusiones y trabajo futuro

En este artículo se presenta el desarrollo de una interfaz conversacional que permite la comunicación e interacción entre un usuario y un actor virtual a través de la voz.

Los usuarios pueden expresarse en español para realizar consultas, dar órdenes o hablar con el actor virtual. El sistema entiende las peticiones de los usuarios y genera una respuesta adecuada, que puede, o bien ser tratada por otro módulo del sistema o bien ser devuelta directamente al usuario, generada mediante síntesis de voz.

El actor virtual ha sido dotado de estado de ánimo (seis más el neutral) que se expresa a través de la voz, modulando y sintetizando la emoción correspondiente. El estado emocional del avatar se puede ver modificado a lo largo de una conversación o una presentación dependiendo de la actitud que mantenga el usuario con el avatar o del tema de la

conversación. Las respuestas ofrecidas por el actor también pueden variar dependiendo estado de ánimo en que éste se encuentre en ese momento.

Actualmente, el sistema se está utilizando de forma satisfactoria en aplicaciones de diversa naturaleza, como por ejemplo: la realización de presentaciones virtuales, presentaciones de clases y soporte de prácticas docentes, y como sistema de ayuda a un sistema domótico.

En cuanto al trabajo futuro, consideramos de vital importancia la mejora del conocimiento dinámico del sistema de manera tal que sea capaz de aprender, ya que hasta el momento, sólo se almacena la "historia" de cada una de las conversaciones. De esta forma, se pretende que equipar a esta interfaz conversacional con cierta capacidad de razonamiento y deducción, que permita manejar reglas básicas de conocimiento.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Dirección General de Investigación a través del Proyecto Nº TIN2004-07926 y por el Gobierno de Aragón gracias al Convenio Walqa Ref. 2004/04/86 y al Proyecto Nº CTPP02/2006.

Referencias

[1] Artificial Intelligence Foundation, http://www.alicebot.org/

[2] Artificial Intelligence Markup Language (AIML) Version 1.0.1, http://www.alicebot.org/TR/2001/WD-aiml/

[3] Barra R., Montero J.M., Macías-Guarasa J., D'Haro L.F., San-Segundo R., Córdoba R. "Prosodic and segmental rubrics in emotion identification", Proc. ICASSP 2006 IEEE International Conference on Acustics, Speech and Signal Processing.

[4] Bolinger D. "Intonation and its uses, melody and grammar in discourse", London: Edward Arnold, 1989.

[5] Boula de Mareüil P., Celerier P., Toen J. "Generation of Emotions by a Morphing Technique in English, French and Spanish", Proc. Speech Prosody 2002, pp. 187-190, 2002.

[6] Cerezo E., Baldassarri S., Serón F. "The use of interactive animated agents for teaching", Proc. ICIE 2007: 3rd International Conference on Interdisciplinarity in Education. 2007.

[7] Cowie R., Douglas-Cowie E., Shroder M. (eds): "Proceedings of the ICSA Workshop on Speech and Emotion: A Conceptual Framework for Research". Belfast, 2000.

[8] Ekman P. "Facial Expression, The Handbook of Cognition and Emotion" John Wiley and Sons, 1999.

[9] Francisco V., Gervás P., Hervás R. "Expression of emotions in the synthesis of voice in contexts narrative". Proc. UCAmI2005, pp.353-360, 2005.

[10] Hoult Christopher, "Emotion in Speech Synthesis", 2004

[11] Iriondo I., Guaus R., Rodriguez A., Lázaro P., Montoya N., Blanco J. M., Bernadas D., Oliver J. M., Tena D., Longth L. "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques". Proc. ISCA 2000, pp.161-166, 2000.

[12] Loquendo, http://www.loquendo.com/

[13] Microsoft Speech API 5.1 (SAPI5) http://www.microsoft.com/speech/default.mspx

[14] Montero J.M, Gutierrez-Arriola J., Colas J., Enriquez E., Pardo J.M, "Analysis and modelling of emotional speech in Spanish", Proceedings of the 14th International Conference on Phonetic, pp. 957-960, 1999.

[15] Murray I., Arnott J. "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion", Journal of the Acoustical Society of America, Vol. 93 (2), pp. 1097-1108, 1993.

[16] Pantic M., Rothktantz L. "Toward an Affect-Sensitive Multimodal Human-Computer Interaction", Proceedings of the IEEE, Vol. 91 (9), pp. 1370-1390, 2003.

[17] Proyect CyN, http://www.daxtron.com/cyn.htm

[18] Seron F., Baldassarri S., Cerezo E.: "MaxinePPT: Using 3D Virtual Characters for Natural Interaction" Proc. 2nd International Workshop on Ubiquitous Computing & Ambient Intelligence, pp. 241-250, 2006.

[19] Shroder M. “Emotional Speech Synthesis: A review”, Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH’01), Vol. 1, pp. 561-564, 2001.